Sparse Principal Components Analysis

Iain M. Johnstone and Arthur Yu Lu
Stanford University and Renaissance Technologies

January 1, 2004

Extended Abstract

Principal components analysis (PCA) is a classical method for the reduction of dimensionality of data in the form of n observations (or cases) of a vector with p variables. Contemporary data sets often have p comparable to, or even much larger than n. Our main assertions, in such settings, are (a) that some initial reduction in dimensionality is desirable before applying any PCA-type search for principal modes, and (b) the initial reduction in dimensionality is best achieved by working in a basis in which the signals have a sparse representation. We describe a simple asymptotic model in which the estimate of the leading principal component vector via standard PCA is consistent if and only if $p(n)/n \to 0$. We provide a simple algorithm for selecting a subset of coordinates with largest sample variances, and show that if PCA is done on the selected subset, then consistency is recovered, even if $p(n) \gg n$.

Our main setting is that of signals and images, in which the number of sampling points, or pixels, is often comparable with or larger than the number of cases, n. Our particular example here is the electrocardiogram (ECG) signal of the beating heart, but similar approaches have been used, say, for PCA on libraries of face images.

Standard PCA involves an $O(\min(p^3, n^3))$ search for directions of maximum variance. But if we have some *a priori* way of selecting $k \ll \min(n, p)$ coordinates in which most of the variation among cases is to be found, then the complexity of PCA is much reduced, to $O(k^3)$. This is a computational reason, but if there is instrumental or other observational noise in each case that is uncorrelated with or independent of relevant case-to-case variation, then there is another compelling reason to preselect a small subset of variables before running PCA.

Indeed, we construct a model of factor analysis type and show that ordinary PCA can produce a consistent (as $n \to \infty$) estimate of the principal factor if and only if $p(n)$ is asymptotically of smaller order than n. Heuristically, if $p(n) \geq cn$, there is so much observational noise and so many dimensions over which to search, that a spurious noise maximum will always drown out the true factor.

Fortunately, it is often reasonable to expect such small subsets of variables to exist: Much recent research in signal and image analysis has sought orthonormal basis and related systems in which typical signals have *sparse* representations: most co-ordinates have small signal energies. If such a basis is used to represent a signal – we use wavelets as the classical example here – then the variation in many coordinates is likely to be very small.

Consequently, we study a simple "sparse PCA" algorithm with the following ingredients: a) given a suitable orthobasis, compute coefficients for each case, b) compute sample variances (over cases) for each coordinate in the basis, and select the k coordinates of largest sample variance, c) run standard PCA on the selected k coordinates, obtaining up to k estimated eigenvectors, d) if desired, use soft or hard thresholding to denoise these estimated eigenvectors, and e) re-express the (denoised) sparse PCA eigenvector estimates in the original signal domain.

We illustrate the algorithm on some exercise ECG data, and also develop theory to show in a single factor model, under an appropriate sparsity assumption, that it indeed overcomes the inconsistency problems when $p(n) \geq cn$, and yields consistent estimates of the principal factor.

1 Introduction

Suppose $\{x_i, i = 1, \ldots, n\}$ is a dataset of n observations on p variables. Standard principal components analysis (PCA) looks for vectors ξ that maximize

$$\text{Var}\,(\xi^T x_i)/\|\xi\|^2. \tag{1}$$

If $\xi_1, \ldots, \xi_k$ have already been found by this optimization, then the maximum defining ξ_{k+1} is taken over vectors ξ orthogonal to $\xi_1, \ldots, \xi_k$.

Our interest lies in situations in which each x_i is a realization of a possibly high dimensional signal, so that p is comparable in magnitude to n, or may even be larger. In addition, we have in mind settings in which the signals x_i contain localized features, so that the principal modes of variation sought by PCA may well be localized also.

Consider, for example, the sample of an electrocardiogram (ECG) in Figure 1 showing some 13 consecutive heart beat cycles as recorded by one of the standard ECG electrodes. Individual beats are notable for features such as the sharp spike ("QRS complex") and the subsequent lower peak ("T wave"), shown schematically in the second panel. The presence of these local features, of differing spatial scales, suggests the use of wavelet bases for efficient representation. Traditional ECG analysis focuses on averages of a series of beats. If one were to look instead at beat to beat *variation*, one might expect these local features to play a significant role in the principal component eigenvectors.

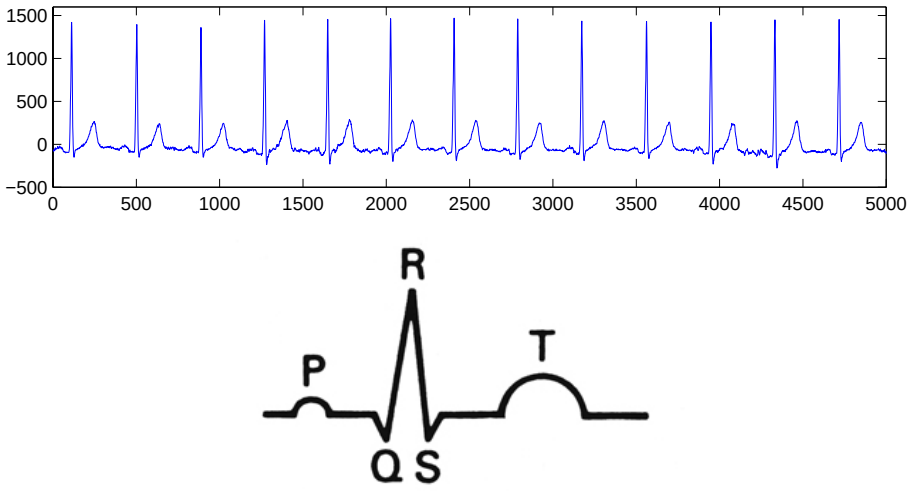

Figure 1: (a) Sample of thirteen beats from one electrode of an electrocardiogram taken in the laboratory of Victor Froelicher, MD, Palo Alto VA. (b) Cartoon of the key features of the cardiac cycle reflected in the ECG trace, from Hampton (1997).

Returning to the general situation, the main contentions of this paper are:

(a) that when p is comparable to n, some reduction in dimensionality is desirable before applying any PCA-type search for principal modes, and

(b) the reduction in dimensionality is best achieved by working in a basis in which the signals have a sparse representation.

We will support these assertions with arguments based on statistical performance and computational cost.

We begin, however, with an illustration of our results on a simple constructed example. Consider a single component (or single factor) model, in which, when viewed as $p-$dimensional column vectors

$$x_i = v_i\rho + \sigma z_i, \qquad i = 1, \ldots, n \tag{2}$$

in which $\rho \in \mathbb{R}^p$ is the single component to be estimated, $v_i \sim N(0,1)$ are i.i.d. Gaussian random effects and $z_i \sim N_p(0, I)$ are independent $p-$dimensional noise vectors.

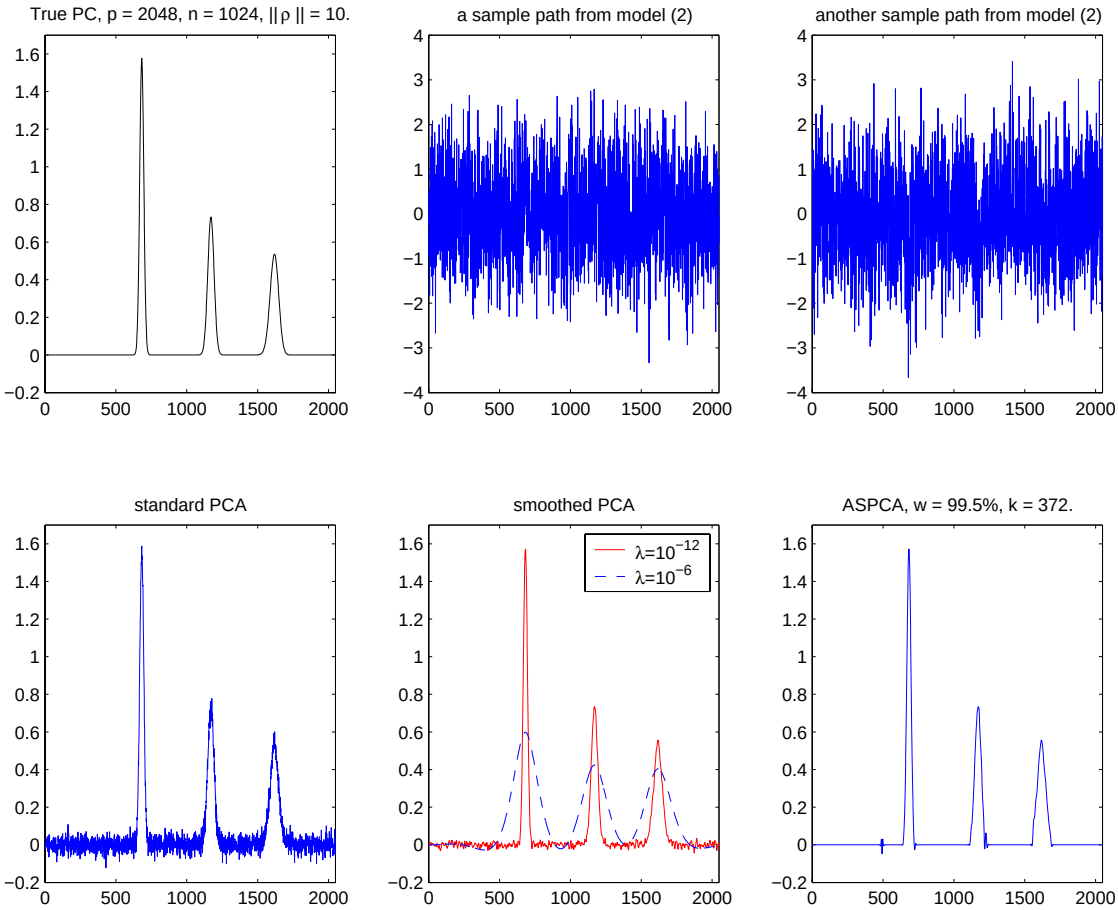

Figure 2: True principal component, the "3-peak" curve. Panel (a): the single component $\rho_l = f(l/n)$ where $f(t) = C\{0.7B(1500, 3000) + 0.5B(1200, 900) + 0.5B(600, 160)\}$ and $B(a, b)(t) = [\Gamma(a + b)/(\Gamma(a)\Gamma(b))]t^{a-1}(1 - t)^{b-1}$ denotes the Beta density on $[0, 1]$. Panels (b,c): Two sample paths drawn from model (2) with $\sigma = 1$. $n = 1024$ replications in total, $p = 2048$. (d): Sample principal component by standard PCA. (e): Sample principal component by smoothed PCA using $\lambda = 10^{-12}$ and $\lambda = 10^{-6}$. (f): Sample principal component by sparse PCA with weighting function $w = 99.5\%$, $k = 372$.

Panel (a) of Figure 2 shows an example of ρ with $p = 2048$ and the vector $\rho_l = f(l/n)$ where $f(t)$ is a mixture of Beta densities on $[0, 1]$, scaled so that $\|\rho\| = (\sum_1^p \rho_l^2)^{1/2} = 10$. Panels (b) and (c) show two sample paths from model (2): the random effect $v_i\rho$ is hard to discern individual cases. Panel (d) shows the result of standard PCA applied to $n = 1024$ observations from (2) with $\sigma = 1$. The effect of the noise remains clearly visible in the estimated principal eigenvector.

For functional data of this type, a regularized approach to PCA has been proposed by Rice & Silverman (1991) and Silverman (1996), see also Ramsay & Silverman (1997) and references therein. While smoothing can be incorporated in various ways, we illustrate the method discussed also in Ramsay & Silverman (1997, Ch. 7), which replaces (1) with

$$\mathrm{Var}\,(\xi^T x_i)/[\|\xi\|^2 + \lambda\|D^2\xi\|^2], \tag{3}$$

where $D^2\xi$ is the $(p-2)\times 1$ vector of second differences of ξ and $\lambda \in (0,\infty)$ is the regularization parameter.

Panel (e) shows the estimated first principal component vector found by maximizing (3) with $\lambda = 10^{-12}$ and $\lambda = 10^{-6}$ respectively. Neither is really satisfactory as an estimate: the first recovers the original peak heights, but fails fully to suppress the remaining baseline noise, while the second grossly oversmooths the peaks in an effort to remove all trace of noise. Further investigation with other choices of λ confirms the impression already conveyed here: no single choice of λ succeeds both in preserving peak heights and in removing baseline noise.

Panel (f) shows the result of the adaptive sparse PCA algorithm to be introduced below: evidently both goals are accomplished quite satisfactorily in this example.

2 The need to select subsets: (in)consistency of classical PCA

A basic element of our sparse PCA proposal is initial selection of a relatively small subset of the initial p variables before any PCA is attempted. In this section, we formulate some (in)consistency results that motivate this initial step.

Consider first the single component model (2). The presence of noise means that the sample covariance matrix $S = n^{-1}\sum_{i=1}^n x_i x_i^T$ will typically have $\min(n,p)$ non-zero eigenvalues. Let $\hat\rho$ be the unit eigenvector associated with the largest sample eigenvalue—with probability one it is uniquely determined up to sign.

One natural measure of the closeness of $\hat\rho$ to ρ uses the angle $\angle(\hat\rho,\rho)$ between the two vectors. We decree that the signs of $\hat\rho$ and ρ be taken so that $\angle(\hat\rho,\rho)$ lies in $[0,\pi/2]$. It will be convenient to phrase the results in terms of an equivalent distance measure

$$\mathrm{dist}(\hat\rho,\rho) = \sin\angle(\hat\rho,\rho) = \sqrt{1-(\rho^T\hat\rho)^2}. \tag{4}$$

For asymptotic results, we will assume that there is a sequence of models (2) indexed by n. Thus, we allow $p(n)$ and $\rho(n)$ to depend by n, though the dependence will usually not be shown explicitly. [Of course σ might also be allowed to vary with n, but for simplicity it is assumed fixed.]

Our first interest is whether the estimate $\hat\rho$ is consistent as $n\to\infty$. This turns out to depend crucially on the limiting value

$$\lim_{n\to\infty} p(n)/n = c. \tag{5}$$

We will also assume that

$$\lim_{n\to\infty} \|\rho(n)\| = \varrho > 0. \tag{6}$$

One setting in which this last assumption may be reasonable is when $p(n)$ grows by adding finer scale wavelet coefficients of a fixed function as n increases.

Theorem 1. *Assume model (2), (5) and (6). Define*

$$\zeta(\tau; c) = \frac{4\sqrt{c}}{\tau}\Big(1 + \frac{2+\sqrt{c}}{\tau}\Big).$$

Then with probability one as $n \to \infty$,

$$\limsup_{n\to\infty} \ \sin \angle(\hat{\rho}, \rho) \le \zeta(\varrho/\sigma, c), \tag{7}$$

so long as the right side is at most one.

For the proof, see Appendix A.2. The bound $\zeta(\tau; c)$ is decreasing in the "signal-to-noise" ratio $\tau = \varrho/\sigma$ and increasing in the dimension-to-sample size ratio $c = \lim p/n$. It approaches 0 as $c \to 0$, and in particular it follows that $\hat{\rho}$ is consistent if $p/n \to 0$.

The proof is based on an almost sure bound for eigenvectors of perturbed symmetric matrices. It appears to give the correct order of convergence: in the case $p/n \to 0$, we have

$$\zeta(\tau, p/n) \sim c(\tau)\sqrt{p/n},$$

with $c(\tau) = 4\tau^{-1} + 8\tau^{-2}$, and examination of the proof shows that in fact

$$\angle(\hat{\rho}, \rho) = O_p(\sqrt{p/n})$$

which is consistent with the $n^{-1/2}$ convergence rate that is typical when p is fixed.

However if $c > 0$, the upper bound (7) is strictly positive. And it turns out that $\hat{\rho}$ must be an inconsistent estimate in this setting:

Theorem 2. *Assume model (2), (5) and (6). If $p/n \to c > 0$, then $\hat{\rho}$ is inconsistent:*

$$\liminf_{n\to\infty} E\angle(\hat{\rho}, \rho) > 0.$$

In short, $\hat{\rho}$ is a consistent estimate of ρ if and only if $p = o(n)$. The noise does not average out if there are too many dimensions p relative to sample size n. A heuristic explanation for this phenomenon is given just before the proof in Appendix A.3.

The inconsistency criterion extends to a considerably more general *multi-component* model. Assume that we have n curves x_i, observed at p time points. Viewed as p dimensional column vectors, this model assumes that

$$x_i = \mu + \sum_{j=1}^{m} v_i^j \rho^j + \sigma z_i, \quad i = 1, \cdots, n. \tag{8}$$

Here μ is the mean function, which is assumed known, and hence is taken to be zero. We make the following assumptions:

(a) The $\rho^j, j = 1, ..., m \le p$ are unknown, mutually orthogonal principal components, with norms $\rho_j(n) = \|\rho^j\|$

$$\|\rho^1\| > \|\rho^2\| \ge \cdots \ge \|\rho^m\|. \tag{9}$$

(b) The multipliers $v_i^j \sim N(0, 1)$ are all independent over $j = 1, \ldots, m$ and $i = 1, \ldots, m$.

(c) The noise vectors $z_i \sim N_p(0, I)$ are independent among themselves and also of the random effects $\{v_i^j\}$.

For asymptotics, we add

(d) We assume that $p(n), m(n)$ and $\{\rho^j(n), j = 1, \ldots, m\}$ are functions of n, though this will generally not be shown explicitly. We assume that the norms of the n^{th} principal components converge as sequences in $\ell_1(\mathbb{N})$:

$$\begin{aligned} \varrho(n) &= (\|\rho^1(n)\|, \ldots, \|\rho^j(n)\|, \ldots) \\ &\to \varrho = (\varrho_1, \ldots, \varrho_j, \ldots). \end{aligned} \tag{10}$$

We write ϱ_+ for the limiting ℓ_1 norm:

$$\varrho_+ = \sum_j \varrho_j.$$

Remark on Notation. The index j, which runs over principal components, will be written as a superscript on *vectors* v^j, ρ^j and u^j(defined in Appendix), but as a subscript on *scalars* such as $\varrho_j(n)$ and ϱ_j.

We continue to focus on the estimation of the principal eigenvector ρ^1, and establish a more general version of the two preceding theorems.

Theorem 3. *Assume model (8) together with conditions (a)-(d). If $p/n \to c$, then*

$$\limsup_{n\to\infty} \sin \angle(\hat{\rho}^1, \rho^1) \le \frac{4\sigma\sqrt{c}}{\varrho_1^2 - \varrho_2^2}[\rho_+ + (2 + \sqrt{c})\sigma]$$

so long as the right side is at most, say, 4/5.

If $c > 0$, then

$$\liminf_{n\to\infty} E\angle(\hat{\rho}^1, \rho^1) > 0.$$

Thus, it continues to be true in the multicomponent model that $\hat{\rho}^1$ is consistent if and only if $p = o(n)$.

3 The sparse PCA algorithm

The inconsistency results of Theorems 2 and 3 emphasize the importance of reducing the number of variables before embarking on PCA, and motivate the sparse PCA algorithm to be described in general terms here. Note that the algorithm *per se* does not require the specification of a particular model, such as (8).

1. *Select Basis.* Select a basis $\{e_\nu\}$ for $\mathbb{R}^p$ and compute co-ordinates $(x_{i\nu})$ for each x_i in this basis:

$$x_i(t) = \sum_\nu x_{i\nu} e_\nu(t), \qquad i = 1, \ldots, n.$$

[The wavelet basis is used in this paper, for reasons discussed in the next subsection.]

2. *Subset.* Calculate the sample variances $\hat{\sigma}_\nu^2 = \widehat{Var}(x_{i\nu})$. Let $\hat{I}$ denote the set of indices ν corresponding to the largest k variances.

[k may be specified in advance, or chosen based on the data, see Section 3.2 below].

3. *Reduced PCA.* Apply standard PCA to the reduced data set $\{x_{i\nu}, \nu \in \hat{I}, i = 1, \ldots, n\}$ on the selected $k-$dimensional subset, obtaining eigenvectors $\hat{\rho}^j = (\hat{\rho}_\nu^j), j = 1, \ldots, k$.

4. *Thresholding.* Filter out noise in the estimated eigenvectors by hard thresholding

$$\hat{\rho}_\nu^{*j} = \eta_H(\hat{\rho}_\nu^j, \delta).$$

[Hard thresholding is given, as usual, by $\eta_H(x,\delta) = xI\{|x| \geq \delta\}$. An alternative is soft thresholding $\eta_S(x,\delta) = \mathrm{sgn}(x)(|x|-\delta)_+$, but hard thresholding has been used here because it preserves the magnitude of retained signals.

The threshold δ can be chosen, for example, by trial and error, or as $\delta = \hat{\tau}_j\sqrt{2\log k}$ for some estimate $\hat{\tau}_j$. In this paper, estimate (13) is used. Another possibility is to set $\hat{\tau}_j = MAD\{\hat{\rho}_\nu^j, \nu = 1,\ldots,k\}/0.6745$.]

5. *Reconstruction.* Return to the original signal domain, setting

$$\hat{\rho}_j(t) = \sum_\nu \hat{\rho}_\nu^{*j} e_\nu(t).$$

In the rest of this section, we amplify on and illustrate various aspects of this algorithm. Given appropriate eigenvalue and eigenvector routines, it is not difficult to code. For example, `MATLAB` files that produce most figures in this paper will soon be available at `www-stat.stanford.edu/~imj/` – to exploit wavelet bases, they make use of the open-source library `WaveLab` available at `www-stat.stanford.edu/~wavelab/`.

3.1 Sparsity and Choice of basis

Suppose that in the basis $\{e_\nu(t)\}$ a population principal component $\rho(t)$ has coefficients $\{\rho_\nu\}$:

$$\rho(t) = \sum_{\nu=1}^{p} \rho_\nu e_\nu(t).$$

It is desirable, both from the point of view of economy of representation, as well as computational complexity, for the expansion in basis $\{e_\nu\}$ to be *sparse*, i.e., most coefficients ρ_ν are small or zero.

One way to formalize this is to require that the ordered coefficient magnitudes decay at some algebraic rate. We say that ρ is contained in a weak ℓ_q ball of radius C, $\rho \in w\ell_q(C)$, if $|\rho|_{(1)} \geq |\rho|_{(2)} \geq \ldots$ and

$$|\rho|_{(\nu)} \leq C\nu^{-1/q}, \qquad \nu = 1,2,\ldots$$

Wavelet bases typically provide sparse representations of one-dimensional functions that are smooth or have isolated singularities or transient features, such as in our ECG example. Here is one such result. Expand ρ in a nice wavelet basis $\{\psi_{jk}(t)\}$ to obtain $\rho = \sum_{jk}\rho_{jk}\psi_{jk}(t)$ and then order coefficients by absolute magnitude, so that (ρ_ν) is a re-ordering of the $|\rho_{jk}|$ in decreasing order. Then smoothness (as measured by membership in some Besov space $B^\alpha_{p,q}$) implies sparsity in the sense that

$$\rho \in B^\alpha_{p,q} \quad \Rightarrow \quad (\rho_\nu) \in w\ell_p, \qquad p = 2/(2\alpha+1).$$

[for details, see Donoho (1993) and Johnstone (2002): in particular it is assumed that $\alpha > (1/p - 1/2)_+$ and that the wavelet ψ is sufficiently smooth.]

In this paper, we will assume that the basis $\{e_\nu\}$ is fixed in advance – and it will generally be taken to be a wavelet basis. Extension of our results to incorporate basis selection (e.g. from a library of orthonormal bases such as wavelet packets) is a natural topic for further research.

3.2 Adaptive choice of k

Here are two possibilities for adaptive choice of $\hat{k} = |\hat{I}|$ from the data:

(a) choose co-ordinates with variance exceeding the estimated noise level by a specified fraction α_n:

$$\hat{I} = \{\nu \ : \ \hat{\sigma}_\nu^2 \geq \hat{\sigma}^2(1+\alpha_n)\}.$$

This choice is considered further in Section 3.5.

(b) As motivation, recall that we hope that the selected set of variables $\hat{I}$ is both small in cardinality and also captures most of the variance of the population principal components, in the sense that the ratio

$$\sum_{\nu \in \hat{I}} \rho_\nu^2 \Big/ \sum_\nu \rho_\nu^2$$

is close to one for the leading population principal components in $\{\rho^1, \ldots, \rho^m\}$. Now let $\chi^2_{(n),\alpha}$ denote the upper $\alpha-$percentile of the $\chi^2_{(n)}$ distribution – if all co-ordinates were pure noise, one might expect $\hat{\sigma}^2_{(\nu)}$ to be close to $n^{-1}\hat{\sigma}^2\chi^2_{(n),\nu/n}$. Define the excess over these percentiles by

$$\hat{\tau}^2_{(\nu)} = \max\{\hat{\sigma}^2_{(\nu)} - n^{-1}\hat{\sigma}^2\chi^2_{(n),\nu/n}, 0\},$$

and for a specified fraction $w(n)$, set

$$\hat{I} = \{\nu \ : \ \sum_{\nu=1}^{\hat{k}} \hat{\tau}^2_{(\nu)} \geq w(n) \sum_\nu \hat{\tau}^2_{(\nu)}\},$$

where $\hat{k}$ is the smallest index k for which the inequality holds. This second method has been used for the figures in this paper, typically with $w(n) = .995$.

Estimation of σ. If the population principal components ρ^j have a sparse representation in basis $\{e_\nu\}$, then we may expect that in most co-ordinates ν, $\{x_{i\nu}\}$ will consist largely of noise. This suggests a simple estimate of the noise level on the assumption that the noise level is the same in all co-ordinates, namely

$$\hat{\sigma}^2 = \text{median}(\hat{\sigma}^2_\nu). \tag{11}$$

3.3 Computational complexity

It is straightforward to estimate the cost of sparse PCA by examining its main steps:

1. This depends on the choice of basis. In the wavelet case no more than $O(np\log p)$ operations are needed.
2. Sort the sample variances and select $\hat{I}$: $O(p\log p)$.
3. Eigendecomposition for a $k \times k$ matrix: $O(k^3)$.
4. Estimate $\hat{\sigma}^2$ and $\widehat{\|\rho\|^2}$: $O(p)$.
5. Apply thresholding: $O(k)$.
6. Reconstruct eigenvectors in the original sample space: $O(k^2p)$.

Hence, the total cost of sparse PCA is

$$O(np \log p + k^2 p).$$

Both standard and smoothed PCA need at least $O((p \wedge n)^3)$ operations. Therefore, if we can find a sparse basis such that $k/p \to 0$, then under the assumption that $p/n \to c$ as $n \to \infty$,the total cost of sparse PCA is $o(p^3)$. We will see in examples to follow that the savings can be substantial.

3.4 Simulated examples

The two examples in this section are both motivated by functional data with localized features.

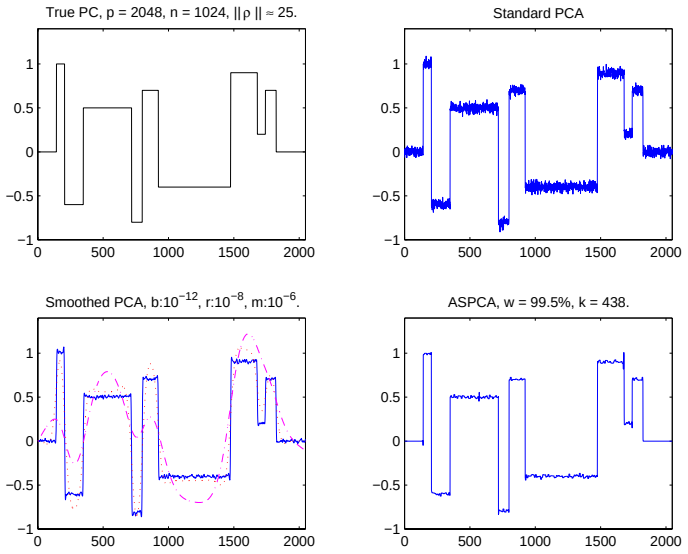

Figure 3: Comparison of the sample principal components for a step function. (a) True principal component $\rho_l = f(l/n)$, the "step" function (b): Sample principal component by standard PCA. (c): Sample principal component by smoothed PCA using $\lambda = 10^{-12}, 10^{-8}$ and 10^{-6}. (d): Sample principal component by sparse PCA with weighting function $w = 99.5\%$, $k = 438$.

The first is a three-peak principal component depicted in Figure 2, and already discussed in Section 1. The second example, Figure 3, has an underlying first principal component composed of step functions. For both examples, the dimension of data vectors is $p = 2048$, the number of observations $n = 1024$, and the noise level $\sigma = 1$. However, the amplitudes of ρ differ, with $\|\rho\| = 10$ for the "3-peak" function and $\|\rho\| \approx 25$ for the "step" function.

Panels (d) and (b) in the two figures respectively show the sample principal components obtained by using standard PCA. While standard PCA does capture the peaks and steps, it retains significant noise in the flat regions of the function. Corresponding panels (e) and (c) show results from smooth PCA with the indicated values of the smoothing parameter. Just as for the three peak curve discussed earlier, in the case of the step function, none of the three estimates simultaneously captures both jumps and flat regions well.

Panels (f) and (d) present the principal components obtained by sparse PCA. Using method (b) of the previous section with $w = 99.5\%$, the *Subset* step selects $k = 372$ and 438 for the "3-peak" curve and "step" function, respectively. The sample principal component

	Standard PCA	Smoothed $\lambda : 10^{-12}$	Smoothed $\lambda : 10^{-6}$	Sparse PCA
ASE (3-peak)	9.681e-04	1.327e-04	3.627e-2	7.500e-05
Time (3-peak)	$\sim$ 12min	$\sim$ 47 min	$\sim$ 43 min	1 min 15 s
ASE (step)	9.715e-04	3.174e-3	1.694e-2	1.947e-04
Time (step)	$\sim$ 12min	$\sim$ 47 min	$\sim$ 46 min	1 min 31 s

Table 1: Accuracy and efficiency comparison

in Figure 2(d) is clearly superior to the other sample p.c.s in Figure 2. Although the principal component function in the step case appears to be only slightly better than the solid blue smooth PCA estimate, we will see later that its squared error is reduced by more than 90%.

Table 1 compares the accuracy of the three PCA algorithms, using average squared error (ASE) defined as

$$\text{ASE} = p^{-1}\|\hat{\rho} - \rho\|^2.$$

The average ASE over 50 iterations is shown. The running time is the CPU time for a single iteration used by Matlab on a MIPS R10000 195.0MHz server.

Figure 4 presents box plots of ASE for the 50 iterations. Sparse PCA gives the best result for the "step" curve. For the "3-peak" function, in only a few iterations does sparse PCA generate larger error than smoothed PCA with a small $\lambda = 10^{-12}$. On the average, ASE using sparse PCA is superior to the other methods by a large margin. Overall Table 1 and Figure 5 show that sparse PCA leads to the most accurate principal component while using much less CPU time than other PCA algorithms.

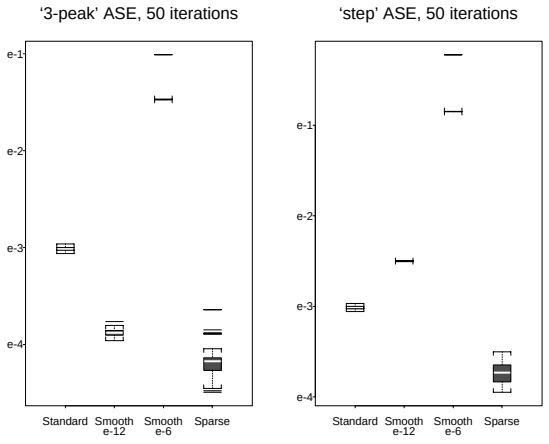

Figure 4: Side-by side box-plots of ASE from 50 iterations using different algorithms. (a) For the "3-peak" function. (b) For the "step" function.

Remarks on the single component model.

Anderson (1963) obtained the asymptotic distribution of $\sqrt{n}(\rho - \hat{\rho})$ for *fixed* p; in particular

$$\mathrm{Var}\{\sqrt{n}(\rho_\nu - \hat{\rho_\nu})\} \to (\|\rho\|^2 + \sigma^2)\frac{\sigma^2}{\|\rho\|^4}(1 - \rho_\nu^2), \tag{12}$$

as $n \to \infty$. For us, p increases with n, but we will nevertheless use (12) as an heuristic basis for estimating the variance $\hat{\tau}$ needed for thresholding. Since the effect of thresholding is to remove noise in small coefficients, setting ρ_ν to 0 in (12) suggests

$$\hat{\tau}_\nu \approx \frac{1}{\sqrt{n}}\frac{\sigma\sqrt{\|\rho\|^2 + \sigma^2}}{\|\rho\|^2}. \tag{13}$$

Neither $\|\rho\|^2$ and σ^2 in (13) are known, but they can be estimated by using the information contained in the sample covariance matrix S, much as in the discussion of Section 3.2. Indeed S_ν^2, the ν-th diagonal element of S, follows a scaled χ^2 distribution, with expectation $\rho_\nu^2 + \sigma^2$. If ρ_ν is a sparse representation of ρ, then most coefficients will be small, suggesting the estimate (11) for σ^2. In the single component model,

$$||\rho||^2 = \sum_1^p \rho_\nu^2 = \sum_1^p E(S_\nu^2) - \sigma^2,$$

which suggests as an estimate:

$$\widehat{||\rho||^2} = \sum_1^p \{S_\nu^2 - \mathrm{median}(S_\nu^2)\}. \tag{14}$$

Figure 5 shows the histograms for these estimates of $\|\rho\|$ and σ based on 100 iterations for the "3-peak" curve and for the "step" function.

3.5 Correct Selection Properties

A basic issue raised by the sparse PCA algorithm is whether the selected subset $\hat{I}$ in fact correctly contains the largest population variances, and only those. We formulate a result, based on large deviations of χ^2 variables, that provides some reassurance.

For this section, assume that the diagonal elements of the sample covariance matrix $S = n^{-1}\sum_1^n x_i x_i^T$ have marginal χ^2 distributions, i.e.,

$$\hat{\sigma}_\nu^2 = S_{\nu\nu} \sim \sigma_\nu^2 \chi_{(n)}^2/n, \qquad \nu = 1, \ldots, p. \tag{15}$$

We will not require any assumptions on the joint distribution of $\{\hat{\sigma}_\nu^2\}$.

Denote the ordered population coordinate variances by $\sigma_{(1)}^2 \geq \sigma_{(2)}^2 \geq \ldots$ and the ordered sample coordinate variances by $\hat{\sigma}_{(1)}^2 \geq \hat{\sigma}_{(2)}^2 \geq \ldots$. A desirable property is that $\hat{I}$ should, for suitable α_n small,
(i) include *all* indices l in

$$I_{in} = \{l : \sigma_l^2 \geq \sigma_{(k)}^2(1 + \alpha_n)\}, \qquad \text{and}$$

(ii) exclude *all* indices l in

$$I_{out} = \{l : \sigma_l^2 \leq \sigma_{(k)}^2(1 - \alpha_n)\}.$$

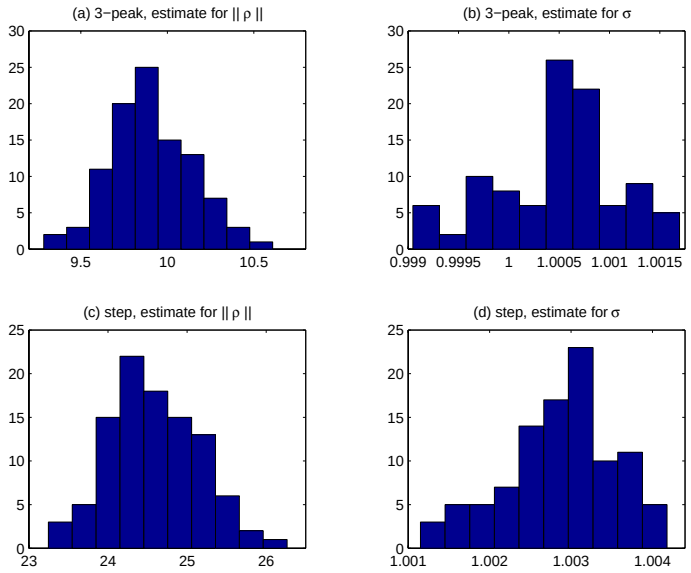

Figure 5: Histograms from 100 iterations. The "3-peak" function, (a) estimate for $\|\rho\| = 10$: mean = 9.91, SD = 0.24. (b): estimate for $\sigma = 1$: mean = 1.0005, SD = .0006. The "step" function, (c): estimate for $\|\rho\| = 24.82$: mean = 24.58, SD = 0.56. (d): estimate for $\sigma = 1$: mean = 1.0029, SD = .0007.

We will show that this in fact occurs if $\alpha_n = \gamma\sqrt{n^{-1}\log n}$, for appropriate $\gamma > 0$.

We say that a *false exclusion* (FE) occurs if any variable in I_{in} is missed:

$$FE = \bigcup_{l \in I_{in}} \{\hat{\sigma}_l^2 < \hat{\sigma}_{(k)}^2\},$$

while a *false inclusion* (FI) happens if any variable in I_{out} is spuriously selected:

$$FI = \bigcup_{l \in I_{out}} \{\hat{\sigma}_l^2 \geq \hat{\sigma}_{(k)}^2\}.$$

Theorem 4. *Under assumptions (15), the chance of an inclusion error of either type in* $\hat{I}_k$ *having magnitude* $\alpha_n = \gamma n^{-1/2}(\log n)^{1/2}$ *is polynomially small:*

$$P\{FE \cup FI\} \leq 2pkn^{-b(\gamma)} + kn^{-(1-2\alpha_n)b(\gamma)},$$

with $b(\gamma) = [\gamma\sqrt{3}/(4 + 2\sqrt{3})]^2$.

For example, if $\gamma = 9$, then $b(\gamma) \doteq 4.36$. As a numerical illustration based on (54) below, if the subset size $k = 50$, while $p = n = 1000$, then the chance of an inclusion error corresponding to a 25% difference in SDs (i.e. $\sqrt{1 + \alpha_n} = 1.25$) is below 5%. That reasonably large sample sizes are needed is a sad fact inherent to variance estimation—as one of Tukey's 'anti-hubrisines' puts it, "it takes 300 observations to estimate a variance to one significant digit of accuracy".

3.6 Consistency

The sparse PCA algorithm is motivated by the idea that if the p.c.'s have a sparse representation in basis $\{e_\nu\}$, then selection of an appropriate subset of variables should overcome the inconsistency problem described by Theorem 2.

To show that such a hope is justified, we establish a consistency result for sparse PCA. For simplicity, we consider the single component model (2), and assume that σ^2 is known—though this latter assumption could be removed by estimating σ^2 using (11).

To select the subset of variables $\hat{I}$, we use a version of rule (a) from Section 3.2:

$$\hat{I} = \{\nu \ : \ \hat{\sigma}_\nu^2 \geq \sigma^2(1+\gamma_n)\}, \tag{16}$$

with $\gamma_n = \gamma(n^{-1}\log n)^{1/2}$ and γ a sufficiently large positive constant—for example $\gamma > \sqrt{12}$ would work for the proof.

We assme that the unknown principal components $\rho = \rho(n)$ satisfy a *uniform sparsity condition*: for some positive constants q, C,

$$\rho(n) \in w\ell_q(C) \quad \text{uniformly in } n. \tag{17}$$

Let $\hat{\rho}_I$ denote the principal eigenvector estimated by step (3) of the sparse PCA algorithm (thresholding is not considered here).

Theorem 5. *Assume that the single component model (2) holds, with $p/n \to c > 0$ and $\|\rho(n)\| \to \varrho > 0$. For each n, assume that $\rho(n)$ satisfies the uniform sparsity condition (17).*

Then the estimated principal eigenvector $\hat{\rho}_I$ obtained by subset selection rule (16) is consistent:

$$\angle(\hat{\rho}_I, \rho) \stackrel{a.s.}{\to} 0.$$

The proof is given in Appendix A.5: it is based on a correct selection property similar to Theorem 4: combined with a modification of the consistency argument for Theorem 3. In fact, the proof shows that consistency holds even under the weaker assumption $p = O(n^a)$, for arbitrary $a > 0$, so long as $\gamma = \gamma(a)$ is set sufficiently large.

3.7 ECG example

This section offers a brief illustration of sparse PCA as applied to some ECG data kindly provided by Jeffrey Froning and Victor Froelicher in the cardiology group at Palo Alto Veterans Affairs Hospital. Beat sequences – typically about 60 cycles in length – were obtained from some 15 normal patients: we have selected two for the preliminary illustrations here.

Data Preprocessing. Considerable preprocessing is routinely done on ECG signals before the beat averages are produced for physician use. Here we describe certain steps taken with our data, in collaboration with Jeff Froning, preparatory to the PCA analysis.

The most important feature of an ECG signal is the Q-R-S complex: the maximum occurs at the R-wave, as depicted in Figure 1(b). Therefore we define the length of one cycle as the gap between two adjacent maxima of R-waves.

1. *Baseline wander* is observed in many ECG data sets, c.f. Figure 6. One common remedy for this problem is to deduct a piecewise linear baseline from the signal, the linear segment (dashed line) between two beats being determined from two adjacent onset points.

The onset positions of R-waves are shown by asterisks. Their exact locations vary for different patients, and as Figure 6 shows, even for adjacent R-waves. The locations are determined manually in this example. To reduce the effect of noise, the values of onset points are calculated by an average of 5 points close to the onset position.

2. Since pulse rates vary even on short time scales, the duration of each heart beat cycle may vary as well. We use linear interpolation to equalize the duration of each cycle,

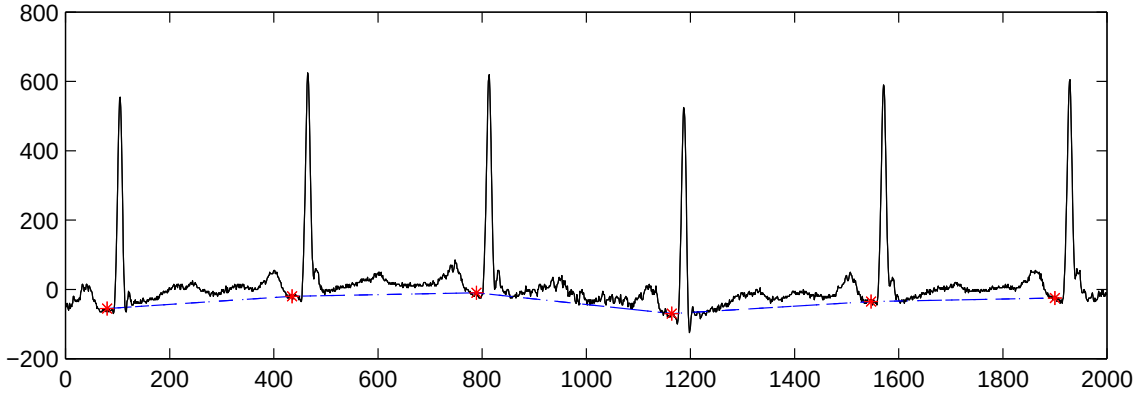

Figure 6: ECG baseline wander.

and for convenience in using wavelet software, discretize to $512 = 2^9$ sample points in each cycle.

3. Finally, due to the importance of the R-wave, the horizontal positions of the maxima are the 150th position in each cycle.

4. Convert the ECG data vector into an $n \times p$ data matrix, where n is the number of observed cycles and $p = 512$. Each row of the matrix presents one heart beat cycle with the maxima of R-waves all aligned at the same position.

PCA analysis. Figure 7 (a) and (d) shows the mean curves for two ECG samples in blue. The number of observations n, i.e. number of heart beats recorded, are 66 and 61, respectively. The first sample principal components for these two sample sets are plotted in plots (c) and (f), with red curves from standard PCA and blue curves from sparse PCA. In both cases there are two sharp peaks in the vicinity of the QRS complex. The first peak occurs shortly before the 150th position, where all the maxima of R-waves are aligned, and the second peak, which has an opposite sign, shortly after.

The standard PCA curve in Figure 6.7.(b, red) is less noisy than that in panel (d, red), even allowing for the difference in vertical scales. Using (11),

$$\hat{\sigma}_1^2 = 24.97 \quad \text{and} \quad \hat{\sigma}_2^2 = 82.12.$$

while the magnitudes of the two mean sample curves are very similar.

The sparse PCA curves (blue) are smoother than the standard PCA ones (red), especially in plot (d) where the signal to noise ratio is lower. On the other hand, the red and blue curves match quite well at the two main peaks. Sparse PCA has reduced noise in the sample principal component in the baseline while keeping the main features.

There is a notable difference between the two estimated p.c.'s. In the first case, the p.c. is concentrated around the R-wave maximum, and the effect is to accelerate or decelerate the rise (and fall) of this peak from baseline in a given cycle. This is more easily seen by comparing plots of $\bar{x} + 2\hat{\rho}$ (green) with $\bar{x} - 2\hat{\rho}$ (red), shown over a magnified part of the cycle in panel (b). In the second case, the bulk of the energy of the p.c. is concentrated in a level shift in the part of the cycle starting with the ST segment. This can be interpreted as beat to beat fluctuation in baseline – since each beat is anchored at 0 at the onset point, there is less fluctuation on the left side of the peak. This is particularly evident in panel

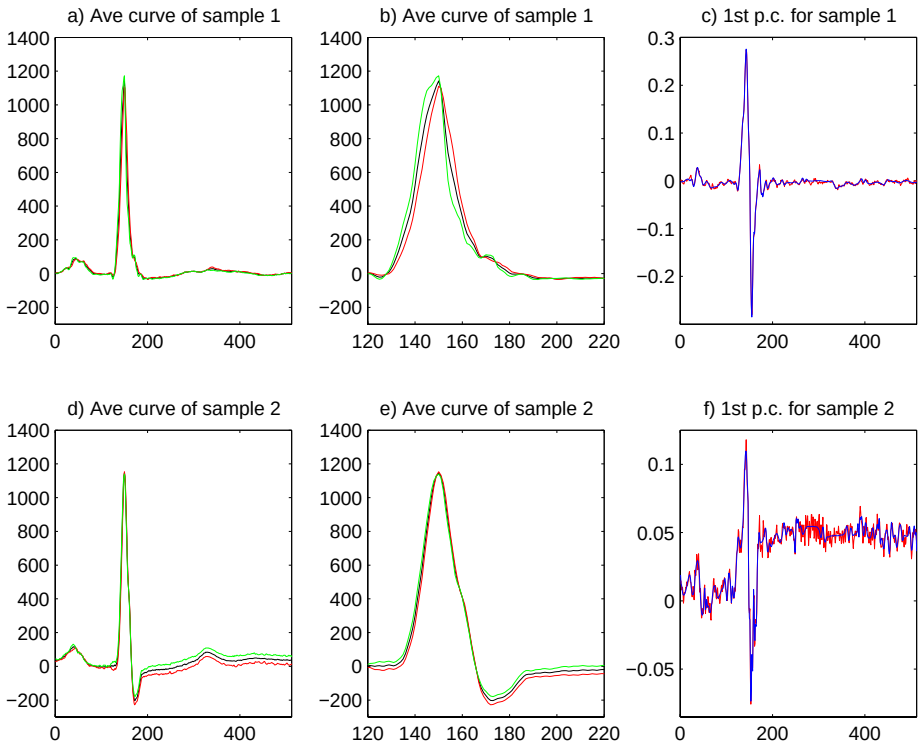

Figure 7: ECG examples. (a): mean curve for ECG sample 1, $n = 66$, in blue, along with $\bar{x} + 2\hat{\rho}$ (green) and $\bar{x} - 2\hat{\rho}$ (red), with $\hat{\rho}$ being the estimated first principal component from sparse PCA (see also (c)). (b) Magnified section of (a) over the range 120-220. (c): First principal components for sample 1 from standard (red) and sparse PCA (blue). (d)– (f): corresponding plots for sample 2, $n = 61$.

(e) – there is again a slight acceleration/deceleration in the rise to the R wave peak – less pronounced in the first case, and also less evident in the fall.

Obvious questions raised by this illustrative example include the nature of effects which may have been introduced by the preprocessing steps, notably the baseline removal anchored at onset points and the alignment of R-wave maxima. Clearly some conventions must be adopted to create rectangular data matrices for p.c. analysis, but detailed analysis of these issues must await future work.

Finally, sparse PCA uses less than 10% of the computing time than standard PCA.

A Appendix

A.1 Preliminaries

Matrices. We first recall some pertinent matrix results. Define the 2−norm of a rectangular matrix by

$$\|A\|_2 = \sup\{\|Ax\|_2 : \|x\|_2 = 1\}. \tag{18}$$

If A is real and symmetric, then $\|A\|_2 = \lambda_{max}(A)$. If $A_{p\times p}$ is partitioned

$$A = \begin{pmatrix} a & b^T \\ b & C \end{pmatrix}$$

where b is $(p-1)\times 1$, then by setting $x = (1\ 0^T)^T$ in (18), one finds that

$$\|b\|_2 \le \|A\|_2. \tag{19}$$

The matrix $B = \rho u^T + u\rho^T$ has at most two non-zero eigenvalues, given by

$$\lambda = (\tau \pm 1)\|\rho\|\|u\|, \qquad \tau = \rho^T u/\|\rho\|\|u\|. \tag{20}$$

Indeed, the identity $\det(I+AC) = \det(I+CA)$ for compatible rectangular matrices A and C means that the non-zero eigenvalues of

$$B = \begin{pmatrix}\rho & u\end{pmatrix}\begin{pmatrix}u^T \\ \rho^T\end{pmatrix}$$

are the same as those of the 2×2 matrix

$$B^* = \begin{pmatrix}u^T \\ \rho^T\end{pmatrix}\begin{pmatrix}\rho & u\end{pmatrix} = \begin{pmatrix}\tau\|\rho\|\|u\| & \|u\|^2 \\ \|\rho\|^2 & \tau\|\rho\|\|u\|\end{pmatrix}$$

from which (20) is immediate.

Angles between vectors. We recall and develop some elementary facts about angles between vectors. The angle between two non-zero vectors ξ, η in $\mathbb{R}^p$ is defined as

$$\angle(\xi,\eta) = \frac{\cos^{-1}|\xi^T\eta|}{\|\xi\|_2\|\eta\|_2} \in [0, \pi/2]. \tag{21}$$

Clearly $\angle(a\xi, b\eta) = \angle(\xi,\eta)$ for non-zero scalars a and b; in fact $\angle(\cdot,\cdot)$ is a metric on one-dimensional subspaces of $\mathbb{R}^p$. If ξ and η are chosen to be unit vectors with $\xi^T\eta \ge 0$, then

$$\|\xi - \eta\|_2 = 2\sin\tfrac{1}{2}\angle(\xi,\eta). \tag{22}$$

The sine rule for plane triangles says that if ξ, η are non-zero and linearly independent vectors in $\mathbb{R}^p$, then

$$\sin\angle(\xi,\eta) = \frac{\|\xi-\eta\|}{\|\xi\|}\sin\angle(\xi-\eta,\eta). \tag{23}$$

These remarks can be used to bound the angle between a vector η and its image under a symmetric matrix M in terms of the angle between η and any principal eigenvector of M.

Lemma 1. *Let ξ be a principal eigenvector of a non-zero symmetric matrix M. For any $\eta \ne 0$,*

$$\angle(\eta, M\eta) \le 3\angle(\eta,\xi).$$

Proof. We may assume without loss of generality that $\|\xi\| = \|\eta\| = 1$ and that $\xi^T\eta \ge 0$. Since ξ is a principal eigenvector of a symmetric matrix, $\|M\xi\| = \|M\|$. From the sine rule (23),

$$\begin{aligned}\sin\angle(M\xi, M\eta) &\le \|M\xi - M\eta\|/\|M\xi\| \\ &\le \|\xi-\eta\| = 2\sin\tfrac{1}{2}\angle(\xi,\eta),\end{aligned}$$

where the final equality uses (22). Some calculus shows that $2\sin\alpha/2 \le \sin 2\alpha$ for $0 \le \alpha \le \pi/4$ and hence

$$\angle(M\xi, M\eta) \le 2\angle(\xi,\eta). \tag{24}$$

From the triangle inequality on angles,

$$\begin{aligned}\angle(\eta, M\eta) &\le \angle(\eta, M\xi) + \angle(M\xi, M\eta) \\ &\le 3\angle(\eta, \xi),\end{aligned}$$

using (24) and the fact that ξ is an eigenvector of M. □

Perturbation bounds. Suppose that a symmetric matrix $A_{p\times p}$ has unit eigenvector q_1. We wish to bound the effect of a *symmetric* perturbation $E_{p\times p}$ on q_1. The following result (Golub & Van Loan (1996, Thm 8.1.10), see also Stewart & Sun (1990)) constructs a unit eigenvector $\hat{q}_1$ of $A+E$ and bounds its distance from q_1 in terms of $\|E\|_2$. Here, the distance between unit eigenvectors q_1 and $\hat{q}_1$ is defined as at (4) and (21).

Let $Q_{p\times p} = [q_1\ Q_2]$ be an orthogonal matrix containing q_1 in the first column, and partition conformally

$$Q^T A Q = \begin{pmatrix} \lambda & 0 \\ 0 & D_{22} \end{pmatrix}, \quad Q^T E Q = \begin{pmatrix} \epsilon & e^T \\ e & E_{22} \end{pmatrix},$$

where D_{22} and E_{22} are both $(p-1)\times(p-1)$.

Suppose that λ is separated from the rest of the spectrum of A; set

$$\delta = \min_{\mu\in\lambda(D_{22})} |\lambda - \mu|.$$

If $\|E\|_2 \le \delta/5$, then there exists $r \in \mathbb{R}^{p-1}$ satisfying

$$\|r\|_2 \le (4/\delta)\|e\|_2 \tag{25}$$

such that

$$\hat{q}_1 = (1 + r^T r)^{-1/2}(q_1 + Q_2 r)$$

is a unit eigenvector of $A+E$. Moreover,

$$\mathrm{dist}(\hat{q}_1, q_1) \le (4/\delta)\|e\|_2.$$

Let us remark that since $\|e\|_2 \le \|E\|_2$ by (19), we have $\|r\|_2 \le 1$ and

$$q_1^T \hat{q}_1 = (1 + \|r\|_2^2)^{-1/2} \ge 1/\sqrt{2}. \tag{26}$$

Suppose now that q_1 is the eigenvector of A associated with the *principal* eigenvalue $\lambda_1(A)$. We verify that, under the preceding conditions, $\hat{q}_1$ is also the principal eigenvector of $A+E$: i.e. if $(A+E)\hat{q}_1 = \lambda^*\hat{q}_1$, then in fact $\lambda^* = \lambda_1(A+E)$.

To show this, we verify that $\lambda^* > \lambda_2(A+E)$. Take inner products with q_1 in the eigenequation for $\hat{q}_1$:

$$\lambda^* q_1^T \hat{q}_1 = q_1^T A \hat{q}_1 + q_1^T E \hat{q}_1. \tag{27}$$

Since A is symmetric, $q_1^T A = \lambda_1(A) q_1^T$. Trivially, we have $q_1^T E \hat{q}_1 \ge -\|E\|_2$. Combine these remarks with (26) to get

$$\lambda^* \ge \lambda_1(A) - \sqrt{2}\|E\|_2.$$

Now $\delta = \lambda_1(A) - \lambda_2(A)$ and since from the minimax characterization of eigenvalues (e.g. Golub & Van Loan (1996, p. 396) or Stewart & Sun (1990, p.218)), $\lambda_2(A+E) \le \lambda_2(A) + \|E\|_2$, we have

$$\begin{aligned}\lambda^* - \lambda_2(A+E) &\ge \delta - (1+\sqrt{2})\|E\|_2 \\ &\ge \delta[1 - (1+\sqrt{2})/5] > 0,\end{aligned}$$

which is the inequality we seek.

Large Deviation Inequalities. If $\bar{X} = n^{-1}\sum_1^n X_i$ is the average of i.i.d. variates with moment generating function $\exp\{\Lambda(\lambda)\} = E\exp\{\lambda X_1\}$, then Cramer's theorem (see e.g. Dembo & Zeitouni (1993, 2.2.2 and 2.2.12)) says that for $x > EX_1$,

$$P\{\bar{X} > x\} \le \exp\{-n\Lambda^*(x)\}, \tag{28}$$

where the conjugate function $\Lambda^*(x) = \sup_\lambda\{\lambda x - \Lambda(\lambda)\}$. The same bound holds for $P\{\bar{X} < x\}$ when $x < EX_1$.

When applied to the $\chi^2_{(n)}$ distribution, with $X_1 = z_1^2$ and $z_1 \sim N(0,1)$, the m.g.f. $\Lambda(\lambda) = -\frac{1}{2}\log(1-2\lambda)$ and the conjugate function $\Lambda^*(x) = \frac{1}{2}[x - 1 - \log x]$. The bounds

$$\log(1+\epsilon) \le \begin{cases} \epsilon - \epsilon^2/2 & -1 < \epsilon < 0, \\ \epsilon - 3\epsilon^2/8 & 0 \le \epsilon < \frac{1}{2}, \end{cases}$$

(the latter following, e.g., from (47) in Johnstone (2001)) yield

$$P\{\chi^2_{(n)} \le n(1-\epsilon)\} \le \exp\{-n\epsilon^2/4\}, \qquad 0 \le \epsilon < 1, \tag{29}$$

$$P\{\chi^2_{(n)} \ge n(1+\epsilon)\} \le \exp\{-3n\epsilon^2/16\}, \quad 0 \le \epsilon < \tfrac{1}{2}. \tag{30}$$

We will use also a slightly sharper bound

$$P\{\chi^2_{(n)} \ge n + t\sqrt{2n}\} \le t^{-1}e^{-t^2/2}. \tag{31}$$

valid for $n \ge 16$ and $0 \le t \le n^{1/6}$ (Johnstone 2001).

When applied to sums of variables $X_1 = z_1 z_2$, with z_1 and z_2 independent $N(0,1)$ variates, the m.g.f. $\Lambda(\lambda) = -\frac{1}{2}\log(1-\lambda^2)$. With $\lambda_*(x) = [(1+4x^2)^{1/2} - 1]/(2x)$, the conjugate function satisfies

$$\Lambda^*(x) = \lambda_* x + \tfrac{1}{2}\log(1-\lambda_*^2) = (3/2)x^2 + O(x^4),$$

as $x \to 0$. Hence, for n large,

$$P\{\bar{X} > \sqrt{bn^{-1}\log n}\} \le Cn^{-3b/2}. \tag{32}$$

Decomposition of sample covariance matrix. Now adopt the multicomponent model (8) along with its assumptions (a) - (c). The sample covariance matrix $S = n^{-1}\sum_1^n x_i x_i^T$ has expectation $ES = R + \sigma^2 I_p$, where

$$R = \sum_{j=1}^m \rho^j \rho^{jT}. \tag{33}$$

Now decompose S according to (8). Introduce $1 \times n$ row vectors $v^{jT} = (v_1^j \cdots v_n^j)$ and collect the noise vectors into a matrix $Z_{p\times n} = [z_1 \cdots z_n]$. We then have

$$S - ES = \sum_{j,k=1}^{m} A^{jk} + \sum_{j=1}^{m} B^j + C. \tag{34}$$

where the $p \times p$ matrices

$$\begin{aligned} A^{jk} &= \Big(n^{-1}\sum_{i=1}^{n} v_i^j v_i^k - \delta_{jk}\Big)\rho^j \rho^{kT} = v_s^{jk}\rho^j\rho^{kT}, \\ B^j &= \sigma n^{-1}\left(\rho^j v^{jT} Z^T + Z v^j \rho^{jT}\right), \\ C &= \sigma^2\big(n^{-1}ZZ^T - I_p\big). \end{aligned} \tag{35}$$

Some limit theorems. We turn to properties of the noise matrix Z appearing in (35). The cross products matrix ZZ^T has a standard p-dimensional Wishart $W_p(n, I)$ distribution with n degrees of freedom and identity covariance matrix, see e.g. Muirhead (1982, p82). Thus the matrix $C = (c_{jk})$ in (34) is simply a scaled and recentered Wishart matrix. We state results below in terms of either ZZ^T or C, depending on the subsequent application. Properties (b) and (c) especially play a key role in inconsistency when $c > 0$.

(a) If $p = O(n)$, then for any $b > 8$,

$$\max_{j,k} |c_{jk}| \le \sigma\sqrt{\frac{b\log n}{n}} \qquad a.s. \qquad \text{as} \quad n \to \infty. \tag{36}$$

Proof. We may clearly take $\sigma = 1$. An off-diagonal term in $n^{-1}ZZ^T = (c_{jk})$ has the distribution of an i.i.d. average $\bar{X} = n^{-1}\sum X_i$ where $X_1 = z_1 z_2$ is the product of two independent standard normal variates. Thus

$$P\{\max_{j\neq k} |c_{jk}| > x\} \le 2p^2 P\{\bar{X} > x\}. \tag{37}$$

Now apply the large deviation bound (32) to the right hand side. Since $p \sim cn$, the Borel-Cantelli lemma suffices to establish (36) for off-diagonal elements for any $b > 2$.

A diagonal term $c_{jj} + 1$ in $n^{-1}ZZ^T$ has the $n^{-1}\chi^2_{(n)}$ distribution. Setting $t = \sqrt{\frac{1}{2}b\log n}$ in (31) yields

$$P\{c_{jj} > \sqrt{bn^{-1}\log n}\} \le \sqrt{2}(b\log n)^{-1/2} n^{-b/4}.$$

Since there are $p \sim cn$ diagonal terms, the conclusion (36) follows (again via Borel-Cantelli) so long as $b > 8$. □

(b) Geman (1980) and Silverstein (1985) respectively established almost sure limits for the largest and smallest eigenvalues of a $W_p(n, I)$ matrix as $p/n \to c \in [0, \infty)$, from which follows:

$$\lambda_1(C), \lambda_p(C) \to \sigma^2(c \pm 2\sqrt{c}). \tag{38}$$

[Although the results in the papers cited are for $c \in (0, \infty)$, the results are easily extended to $c = 0$ by simple coupling arguments.]

(c) Suppose in addition that v is a $1 \times n$ vector with independent $N(0,1)$ entries, which are also independent of Z. Conditioned on v, the vector Zv is distributed as $N_p(0, \|v\|^2 I)$. Since Z is independent of v, we conclude that

$$Zv \stackrel{\mathcal{D}}{=} \chi_{(n)}\chi_{(p)}U_p \tag{39}$$

where $\chi^2_{(n)}$ and $\chi^2_{(p)}$ denote chi-square variables and U_p a vector uniform on the surface of the unit sphere S^{p-1} in $\mathbb{R}^p$, and all three variables are independent.

Now let $u_{p\times 1} = \sigma n^{-1} Zv$. From (39) we have

$$\|u\|^2 \stackrel{\mathcal{D}}{=} \sigma^2 n^{-2}\chi^2_{(n)}\chi^2_{(p)} \stackrel{a.s.}{\to} \sigma^2 c, \tag{40}$$

as $p/n \to c \in [0,\infty)$.

If ρ is any fixed vector in $\mathbb{R}^p$, it follows from (39) that

$$\tau = \tau(p) = \rho^T u/\|\rho\|\|u\| \stackrel{\mathcal{D}}{=} U_{p,1},$$

the distribution of the first component of U_p. It is well known that $U_1^2 \sim \text{Beta}(1/2,(p-1)/2)$, so that $EU_1^2 = p^{-1}$ and $\text{Var}U_1^2 \le 2p^{-2}$. From this it follows that

$$\tau(p) \stackrel{a.s.}{\to} 0, \qquad p \to \infty. \tag{41}$$

(d) Let $u^j = \sigma n^{-1} Z v^j$ be the vectors appearing in the definition of B^j for $1 \le j \le m$. We will show that a.s.

$$\lim_{n\to\infty} \sup_j \|u^j\| < c_0 \tag{42}$$

(the constant $c_0 = 2\sigma(1+\sqrt{c})$ would do).

Proof. Since

$$\|u^j\|^2 = \sigma^2 n^{-2} v^{jT} Z^T Z v^j$$

we have

$$\sup_j \|u^j\|^2 \le \sigma^2 n^{-1} \lambda_{max}(ZZ^T) \sup_j \|v^j\|^2/n. \tag{43}$$

From (38), it follows that w.p. 1, ultimately

$$\lambda_{max}(ZZ^T)/n \le 2(1+\sqrt{c})^2. \tag{44}$$

The squared lengths $\|v^j\|^2$ follow independent $\chi^2_{(n)}$ laws. Since from (28) there exists c_1 for which $P\{\chi^2_{(n)} \ge 2n\} \le e^{-c_1 n}$ for $n \ge n_0$, it follows that

$$P\{\sup \|v^j\|^2/n > 2\} \le p e^{-c_1 n}$$

and so w.p. 1 it is ultimately true that

$$\sup_j \|v^j\|^2/n \le 2. \tag{45}$$

Substituting (44) and (45) into (43), we recover (42). □

A.2 Upper Bounds: Proof of Theorems 1 and 3

Instead of working directly with the sample covariance matrix S, we consider $S^* = S - \sigma^2 I_p$. It is apparent that S^* has the same eigenvectors as S. We decompose $S^* = R + E$, where R is given by (33) and has spectrum

$$\lambda(R) = \{\|\rho^1\|^2, \cdots, \|\rho^m\|^2, 0\}.$$

The perturbation matrix $E = A + B + C$, where A and B refer to the sums in (34).

Proposition 1. *Assume that multicomponent model (8) holds, along with assumptions (a) - (d). For any $\epsilon > 0$, if $p, n \to \infty, p/n \to c$, then almost surely*

$$\limsup \|E\|_2 \leq \sigma\sqrt{c}\sum \varrho_j + \sigma^2(c + 2\sqrt{c}). \tag{46}$$

Proof. We will obtain a bound in the form

$$\|E\|_2 \leq E_n(\omega) = A_n(\omega) + B_n(\omega) + C_n(\omega),$$

where the A_n, B_n and C_n will be given below. We have shown explicitly the dependence on ω to emphasize that these quantities are random. Finally we show that the a.s. limit of $E_n(\omega)$ is the right side of (46).

A term. Introduce symmetric matrices $2\tilde{A}^{jk} = A^{jk} + A^{kj} = v_s^{jk}(\rho^j \rho^{kT} + \rho^k \rho^{jT})$. Since ρ^j and ρ^k are orthogonal, (20) implies that

$$\|\tilde{A}^{jk}\|_2 \leq |v_s^{jk}| \, \|\rho^j\| \, \|\rho^k\|,$$

and so

$$\|\sum_{j,k} A^{jk}\|_2 \leq \max_{j,k} |v_s^{jk}| \Big(\sum_j \|\rho^j\|\Big)^2 =: A_n(\omega).$$

The v_s^{jk} are entries of a scaled and recentered $W_m(n, I)$ matrix, and so by (36), the maximum converges almost surely to 0. Since $\sum_j \|\rho^j\| \to \sum \varrho_j < \infty$, it follows that the A_n-term converges to zero a.s.

B term. Applying (20) to the definition (35) of B^j, we have

$$\|B^j\|_2 \leq X_n(j) = (1 + |\tau_j|)\|\rho^j\|\|u^j\| \overset{a.s.}{\to} \sigma\sqrt{c}\varrho_j$$

where $\tau_j = \rho^{jT} u^j / \|\rho^j\|\|u^j\|$ and $u^j = \sigma n^{-1} Z v^j$, and the convergence follows from (40) and (41).

Since $|\tau_j| \leq 1$ and using (42), we have a.s. that for $n > n(\omega)$,

$$X_n(j) \leq Y_n(j) := 2c_0\|\rho^j\| \to 2c_0\varrho_j.$$

The norm convergence (10) implies that $\sum_j Y_n(j) \to 2c_0 \sum \varrho_j$ and so it follows from the version of the dominated convergence theorem due to Pratt (1960) that

$$\sum \|B^j\|_2 \leq \sum_j X_n(j) =: B_n(\omega) \overset{a.s.}{\to} \sigma\sqrt{c}\sum \varrho_j.$$

C term. Using (38),

$$C_n(\omega) = \|C\|_2 = \lambda_{max}(C) \overset{a.s.}{\to} \sigma^2(c + 2\sqrt{c}).$$

□

Proof of Theorem 3 [Theorem 1 is a special case.] We apply the perturbation theorem with $A = R$ and $E = A + B + C$. The separation between the principal eigenvalue of R and the remaining ones is

$$\delta_n = \rho_1^2(n) - \rho_2^2(n) \to \rho_1^2 - \rho_2^2,$$

while from Proposition 1 we have the bound

$$\|E\|_2 \le E_n(\omega) \overset{a.s.}{\to} \sigma\sqrt{c}\varrho_+ + \sigma^2(c + 2\sqrt{c}).$$

Consequently, if

$$4\sigma\sqrt{c}\varrho_+ + \sigma^2(c + 2\sqrt{c}) \le \varrho_1^2 - \varrho_2^2,$$

then

$$\limsup_{n\to\infty} \ \mathrm{dist}(\hat{\rho}^1, \rho^1) \le \Omega(\rho, c; \sigma),$$

where

$$\Omega(\rho, c; \sigma) = 4\sigma\sqrt{c}[\varrho_+ + \sigma(\sqrt{c} + 2)]/(\varrho_1^2 - \varrho_2^2).$$

A.3 Lower Bounds: Proof of Theorem 2

We begin with a heuristic outline of the proof. We write S in the form $D + B$, introducing

$$D = (1 + v_s)\rho\rho^T + \sigma^2 n^{-1} Z Z^T,$$

while, as before, $B = \rho u^T + u\rho^T$ and $u = \sigma n^{-1} Z v$.

A symmetry trick plays a major role: write $S_- = D - B$ and let $\hat{\rho}_-$ be the principal unit eigenvector for S_-.

The argument makes precise the following chain of remarks, which are made plausible by reference to Figure 8.

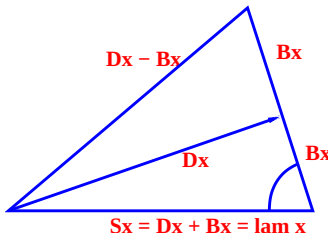

Figure 8: `Needs caption`, with $x \leftarrow \hat{\rho}, lam \leftarrow \hat{\lambda}$

(i) $B\hat{\rho}$ is nearly orthogonal to $D\hat{\rho} + B\hat{\rho} = S\hat{\rho} = \hat{\lambda}\hat{\rho}$.

(ii) the side length $\|B\hat{\rho}\|$ is bounded away from zero, when $c > 0$.

(iii) the angle between $\hat{\rho}$ and $S_-\hat{\rho}$ is "large", i.e. bounded away from zero.

(iv) the angle between $\hat{\rho}$ and $\hat{\rho}_-$ is "large" [this follows from Lemma 1 applied to $M = S_-$.]

(v) and finally, the angle between $\hat{\rho}$ and ρ must be "large", due to the equality in distribution of $\hat{\rho}$ and $\hat{\rho}_-$.

Getting down to details, we will establish (i)-(iii) under the assumption that $\hat{\rho}$ is close to ρ. Specifically, we show that given $\delta > 0$ small, there exists $\alpha(\delta) = \alpha(\delta; \sigma, c) > 0$ such that w.p. $\to 1$,

$$\angle(\hat{\rho}, \rho) \le \delta \quad \Rightarrow \quad \angle(\hat{\rho}, S_-\hat{\rho}) \ge \alpha(\delta).$$

Let $N_\delta = \{x \in \mathbb{R}^p : \angle(x, \rho) \le \delta\}$ be the (two-sided) cone of vectors making angle at most δ with x. We show that on N_δ, both

(ii') $\|Bx\|$ is bounded below (see (49)), and

(i') Bx is nearly orthogonal to x (see (50).

For convenience in this proof, we may take $\|\rho\| = 1$. Write $x \in N_{\delta_1}$ in the form

$$x = (\cos\delta)\rho + (\sin\delta)\eta, \qquad \eta \perp \rho, \|\eta\| = 1, 0 \le \delta \le \delta_1. \tag{47}$$

Since $B\rho = (u^T\rho)\rho + u$ and $B\eta = (u^T\eta)\rho$, we find that

$$Bx = (\cos\delta)u + [(\cos\delta)(u^T\rho) + (\sin\delta)(u^T\eta)]\rho. \tag{48}$$

Denote the second right side term by r: clearly $\|r\| \le |u^T\rho| + (\sin\delta)\|u\|$, and so, uniformly on N_δ,

$$\|Bx\| \ge (\cos\delta - \sin\delta)\|u\| - |u^T\rho|.$$

Since both $\|u\| \to \sigma\sqrt{c}$ and $u^T\rho \to 0$ a.s., we conclude that w.p. $\to 1$,

$$\inf_{N_\delta} \|Bx\| \ge \tfrac{1}{2}\sigma\sqrt{c}\cos\delta. \tag{49}$$

Turning to the angle between x and Bx, we find from (47) and (48) that

$$x^TBx = 2(\cos^2\delta)(\rho^Tu) + 2\cos\delta\sin\delta(u^T\eta),$$

and so, uniformly over N_δ,

$$|x^TBx| \le 2\cos^2\delta|\rho^Tu| + (\sin 2\delta)\|u\|.$$

Consequently, using $\|x\| = 1$ and (49), w.p. $\to 1$, and for $\delta < \pi/4$, say,

$$|\cos\angle(Bx, x)| = \frac{|x^TBx|}{\|x\|\|Bx\|} \le \frac{2\sigma\sqrt{c}\sin 2\delta}{\frac{1}{2}\sigma\sqrt{c}\cos\delta} \le c_2\delta. \tag{50}$$

Now return to Figure 8. As a prelude to step (iii), we establish a lower bound for $\alpha = \angle(\hat{\rho}, D\hat{\rho})$. Applying the sine rule (23) to $\xi = D\hat{\rho}$ and $\eta = \hat{\lambda}\hat{\rho} = D\hat{\rho} + B\hat{\rho}$, we obtain

$$\sin\angle(D\hat{\rho}, \hat{\rho}) = \frac{\|B\hat{\rho}\|}{\|D\hat{\rho}\|}\sin\angle(B\hat{\rho}, \hat{\rho}). \tag{51}$$

On the assumption that $\hat{\rho} \in N_\delta$, bound (50) yields

$$\sin\angle(B\hat{\rho}, \hat{\rho}) \ge \sin(\pi/2 - c_3\delta),$$

and (49) implies that

$$\|B\hat{\rho}\| \ge \tfrac{1}{2}\sigma\sqrt{c}\cos\delta.$$

On the other hand, since $\|\hat{\rho}\| = 1$,

$$\begin{aligned}\|D\hat{\rho}\| \le \|D\| &\le 1 + v_s + \sigma^2\lambda_{max}(n^{-1}ZZ^T) \\ &\le [1 + \sigma^2(1 + \sqrt{c})^2](1 + o(1)),\end{aligned}$$

w.p. 1 for large n.

Combining the last three bounds into (51) shows that there exists a positive $\alpha(\delta;\sigma,c)$ such that if $\hat{\rho} \in N_\delta$, then w.p. 1 for large n,

$$\sin\alpha \geq \sin\alpha(\delta;\sigma,c) > 0.$$

Returning to Figure 8, consider $\angle(D\hat{\rho}+B\hat{\rho}, D\hat{\rho}-B\hat{\rho}) = \alpha+\gamma$. Since $\beta \geq \pi/2 - c_3\delta$, we clearly have $\alpha+\gamma \leq \pi-\beta \leq \pi/2 + c_3\delta$ and hence

$$\begin{aligned}\angle(D\hat{\rho}+B\hat{\rho}, D\hat{\rho}-B\hat{\rho}) &= \min\{\alpha+\gamma, \pi-\alpha-\gamma\} \\ &\geq \min\{\alpha, \pi/2-c_3\delta\}.\end{aligned}$$

In particular, with $\delta \leq \delta_0(\sigma,c)$,

$$\angle(\hat{\rho}, S_-\hat{\rho}) \geq \min\{\alpha(\delta), \pi/2-c_3\delta\} = \alpha(\delta),$$

which is our step (iii). As mentioned earlier, Lemma 1 applied to $M = S_-$ entails that $\angle(\hat{\rho},\hat{\rho}_-) \geq (1/3)\alpha(\delta)$. For the rest of the proof, we write $\hat{\rho}_+$ for $\hat{\rho}$. To summarize to this point, we have shown that if $\angle(\hat{\rho}_+,\rho) \leq \delta$, then w.p. $\to 1$,

$$\angle(\hat{\rho}_+,\hat{\rho}_-) \geq (1/3)\alpha(\delta). \tag{52}$$

Note that S and S_- have the same distribution: viewed as functions of random terms Z and v:

$$S_-(Z,v) = S_+(Z,-v).$$

We call an event $\mathcal{A}$ symmetric if $(Z,v) \in \mathcal{A}$ iff $(Z,-v) \in \mathcal{A}$. For such symmetric events

$$E[\angle(\hat{\rho}_+,\rho),\mathcal{A}] = E[\angle(\hat{\rho}_-,\rho),\mathcal{A}].$$

From this and the triangle inequality for angles

$$\angle(\hat{\rho}_+,\rho) + \angle(\rho,\hat{\rho}_-) \geq \angle(\hat{\rho}_+,\hat{\rho}_-),$$

it follows that

$$E[\angle(\hat{\rho}_+,\rho),\mathcal{A}] \geq \tfrac{1}{2}E[\angle(\hat{\rho}_+,\hat{\rho}_-),\mathcal{A}] \tag{53}$$

Hence

$$E[\angle(\hat{\rho}_+,\rho)] \geq E[\angle(\hat{\rho}_+,\rho),\mathcal{A}^c] + \tfrac{1}{2}E[\angle(\hat{\rho}_+,\hat{\rho}_-),\mathcal{A}].$$

By the symmetry of the distributions, conclusion (52) is also obtained w.p. $\to 1$ if $\angle(\hat{\rho}_-,\rho) \leq \delta$. Consequently, letting $\mathcal{A}$ refer to the symmetric event $\mathcal{A}_\delta = \{\angle(\hat{\rho}_+,\rho) \leq \delta\} \cup \{\angle(\hat{\rho}_-,\rho) \leq \delta\}$, we have

$$\begin{aligned}E[\angle(\hat{\rho}_+,\rho)] &\geq \delta P(\mathcal{A}_\delta^c) + \tfrac{1}{2}E[\angle(\hat{\rho}_+,\hat{\rho}_-),\mathcal{A}_\delta] \\ &\geq \min\{\delta, \alpha(\delta)/6\}(1+o(1)).\end{aligned}$$

This completes the proof of Theorem 2. The lower bound proof for Theorem 3 proceeds similarly, but is omitted – for some extra detail, see Lu (2002).

A.4 Proof of Theorem 4.

We may assume, without loss of generality, that $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_p^2$.

False inclusion. For any fixed constant t,

$$\hat{\sigma}_i^2 \geq t \text{ for } i = 1, \ldots, k \text{ and } \hat{\sigma}_l^2 < t \;\; \Rightarrow \;\; \hat{\sigma}_l^2 < \hat{\sigma}_{(k)}^2.$$

This threshold device leads to bounds on error probabilities using only marginal distributions. For example, consider false inclusion of variable l:

$$P\{\hat{\sigma}_l^2 \geq \hat{\sigma}_{(k)}^2\} \leq \sum_{i=1}^{k} P\{\hat{\sigma}_i^2 < t\} + P\{\hat{\sigma}_l^2 \geq t\}.$$

Write $\bar{M}_n$ for a $\chi_{(n)}^2/n$ variate, and note from (15) that $\hat{\sigma}_\nu^2 \sim \sigma_\nu^2 \bar{M}_n$. Set $t = \sigma_k^2(1-\epsilon_n)$ for a value of ϵ_n to be determined. Since $\sigma_i^2 \geq \sigma_k^2$ and $\sigma_l^2 \leq \sigma_k^2(1-\alpha_n)$, we arrive at

$$\begin{aligned} P\{\hat{\sigma}_l^2 \geq \hat{\sigma}_{(k)}^2\} &\leq kP\{\bar{M}_n < 1-\epsilon_n\} + P\Big\{\bar{M}_n \geq \frac{1-\epsilon_n}{1-\alpha_n}\Big\} \\ &\leq k\exp\Big\{-\frac{n\epsilon_n^2}{4}\Big\} + \exp\Big\{-\frac{3n}{16}\Big(\frac{\alpha_n-\epsilon_n}{1-\alpha_n}\Big)^2\Big\} \end{aligned}$$

using large deviation bound (29). With the choice $\epsilon_n = \sqrt{3}\alpha_n/(2+\sqrt{3})$, both exponents are bounded above by $-b(\gamma)\log n$, and so $P\{FI\} \leq p(k+1)n^{-b(\gamma)}$.

False exclusion. The argument is similar, starting with the remark that for any fixed t,

$$\hat{\sigma}_i^2 \leq t \text{ for } i \geq k, i \neq l \text{ and } \hat{\sigma}_l^2 \geq t \;\; \Rightarrow \;\; \hat{\sigma}_l^2 \geq \hat{\sigma}_{(k)}^2.$$

Consequently, if we set $t = \sigma_k^2(1+\epsilon_n)$ and use $\sigma_l^2 \geq \sigma_k^2(1+\alpha_n)$, we get

$$\begin{aligned} P\{\hat{\sigma}_l^2 < \hat{\sigma}_{(k)}^2\} &\leq \sum_{i \geq k} P\{\hat{\sigma}_i^2 > t\} + P\{\hat{\sigma}_l^2 < t\} \\ &\leq (p-1)P\{\bar{M}_n > 1+\epsilon_n\} + P\big\{\bar{M}_n > \frac{1+\epsilon_n}{1+\alpha_n}\big\} \\ &\leq (p-1)\exp\Big\{-\frac{3n\epsilon_n^2}{16}\Big\} + \exp\Big\{-\frac{n}{4}\Big(\frac{\alpha_n-\epsilon_n}{1+\alpha_n}\Big)^2\Big\}, \end{aligned}$$

this time using (30).

The bound $P\{FE\} \leq pkn^{-b(\gamma)} + ke^{-b(\gamma)(1-2\alpha_n)\log n}$ follows on setting $\epsilon_n = 2\alpha_n/(2+\sqrt{3})$ and noting that $(1+\alpha_n)^{-2} \geq 1-2\alpha_n$.

For numerical bounds, we may collect the preceding bounds in the form

$$P(FE \cup FI) \leq [pk + (p-1)(k-1)]e^{-b(\gamma)\log n} + pe^{-b(\gamma)\log n/(1-\alpha_n)^2} + (k-1)e^{-b(\gamma)\log n/(1+\alpha_n)^2}. \tag{54}$$

A.5 Proof of Theorem 5

Outline. Recall that $\gamma_n = \gamma(n^{-1}\log n)^{1/2}$, that the selected subset of variables $\hat{I}$ is defined by

$$\hat{I} = \{\nu : \hat{\sigma}_\nu^2 \geq \sigma^2(1+\gamma_n)\}$$

and that the estimated principal eigenvector based on $\hat{I}$ is written $\hat{\rho}_I$. We set

$$\rho_I = (\rho_\nu : \nu \in \hat{I}),$$

and will use the triangle inequality $d(\hat{\rho}_I, \rho) \leq d(\hat{\rho}_I, \rho_I) + d(\rho_I, \rho)$ to show that $\hat{\rho}_I \to \rho$. There are three main steps.

(i) Construct deterministic sets of indices

$$I_n^{\pm} = \{\nu : \rho_\nu^2 \geq \sigma^2 a_{\mp} \gamma_n\}$$

which bracket $\hat{I}$ almost surely as $n \to \infty$:

$$I_n^- \subset \hat{I} \subset I_n^+ \qquad \text{w.p. 1.} \tag{55}$$

(ii) the uniform sparsity, combined with $\hat{I}^c \subset I_n^{-c}$, is used to show that

$$d(\rho_I, \rho) \stackrel{a.s.}{\to} 0.$$

(iii) the containment $\hat{I} \subset I_n^+$, combined with $|I_n^+| = o(n)$ shows via methods similar to Theorem 4 that

$$d(\hat{\rho}_I, \rho_I) \stackrel{a.s.}{\to} 0.$$

Details. Step (i). We first obtain a bound on the cardinality of $I_n^{\pm}$ using the uniform sparsity conditions (17). Since $|\rho|_{(\nu)} \leq C\nu^{-1/q}$

$$\begin{aligned} |I_n^{\pm}| &\leq |\{\nu \ : \ C^2\nu^{-2/q} \geq \sigma^2 a_{\mp}\gamma_n\}|, \\ &\leq C^q/(\sigma^2 a_{\mp}\gamma_n)^{q/2} = o(n^{1/2}). \end{aligned}$$

Turning to the bracketing relations (55), we first remark that $\hat{\sigma}_\nu^2 \stackrel{\mathcal{D}}{=} \sigma_\nu^2 \chi_{(n)}^2/n$, and when $\nu \in I_n^{\pm}$,

$$\sigma_\nu^2 = \sigma^2(1 + \rho_\nu^2/\sigma^2) \geq \sigma^2(1 + a_{\mp}\gamma_n).$$

Using the definitions of $\hat{I}$ and writing $\bar{M}_n$ for a random variable with the distribution of $\chi_{(n)}^2/n$, we have

$$\begin{aligned} P_n^- = P(I_n^- \not\subseteq \hat{I}) &\leq \sum_{\nu \in I_n^-} P\{\hat{\sigma}_\nu^2 < \sigma^2(1+\gamma_n)\} \\ &\leq |I_n^-| P\{\bar{M}_n < (1+\gamma_n)/(1+a_+\gamma_n)\}. \end{aligned}$$

We apply (29) with $\epsilon_n = (a_+ - 1)\gamma_n/(1 + a_+\gamma_n)$ and for n large and γ' slightly smaller than γ^2,

$$n\epsilon_n^2 > (a_+ - 1)^2 \gamma' \log n,$$

so that

$$P_n^- \leq cn^{1/2} \exp\{-n\epsilon_n^2/4\} \leq cn^{1/2 - \gamma_+''}$$

with $\gamma_+'' = (a_+ - 1)^2\gamma'/4$. If $\sqrt{\gamma} \geq 12$, then $\gamma_+'' \geq 3$ for suitable $a_+ > 2$.

The argument for the other inclusion is analogous:

$$\begin{aligned} P_n^+ = P(\hat{I} \not\subseteq I_n^+) &\leq \sum_{\nu \notin I_n^+} P\{\hat{\sigma}_\nu^2 \geq \sigma^2(1+\gamma_n)\} \\ &\leq pP\{\bar{M}_n \geq (1+\gamma_n)/(1+a_-\gamma_n)\} \\ &\leq pn^{-\gamma_-''}, \end{aligned}$$

with $\gamma_-'' = 3(1-a_-)^2\gamma'/16$ so long as n is large enough. If $\sqrt{\gamma} \geq 12$, then $\gamma_-'' > 2$ for suitable $a_- < 1 - \sqrt{8/9}$.

By a Borel-Cantelli argument, (55) follows from the bounds on P_n^- and P_n^+.

Step (ii). For $n > n(\omega)$ we have $I_n^- \subset \hat{I}$ and so

$$\|\rho_I - \rho\|^2 = \sum_{\nu \notin \hat{I}} \rho_\nu^2 \leq \sum_{I_n^{-c}} \rho_\nu^2.$$

When $\nu \in I_n^{-c}$, we have by definition

$$\rho_\nu^2(n) < \sigma^2 a_+ \gamma \sqrt{n^{-1}\log n} := \epsilon_n^2,$$

say, while the uniform sparsity condition entails

$$|\rho|_{(\nu)}^2 \leq C^2 \nu^{-2/q}.$$

Putting these together, and defining $s_* = s_*(n)$ as the solution of the equation $Cs^{-1/q} = \epsilon_n$, we obtain

$$\begin{aligned}
\sum_{I_n^{-c}} \rho_\nu^2 &\leq \sum_\nu \epsilon_n^2 \wedge \rho_\nu^2 \\
&= \sum_\nu \epsilon_n^2 \wedge |\rho|_{(\nu)}^2 \\
&\leq \sum_\nu \epsilon_n^2 \wedge C^2 \nu^{-2/q} \\
&\leq \int_0^\infty \epsilon_n^2 \wedge C^2 s^{-2/q} ds \\
&= s_* \epsilon_n^2 + q(2-q)^{-1} C^2 s_*^{1-2/q} \\
&= [2/(2-q)] C^q \epsilon_n^{2-q} \to 0
\end{aligned}$$

as $n \to \infty$.

Step (iii). We adopt the abbreviations

$$\begin{aligned}
u_I &= (u_\nu \ : \ \nu \in \hat{I}), \\
Z_I &= (z_{\nu i} \ : \ \nu \in \hat{I}, i = 1, \ldots, n), \\
S_I &= (S_{\nu\nu'} \ : \ \nu, \nu' \in \hat{I}).
\end{aligned}$$

As in the proof of Theorem 3, we consider $S_I^* = S_I - \sigma^2 I_{\hat{k}} = \rho_I \rho_I^T + E_I$ and note that the perturbation term has the decomposition

$$E_I = v_s \rho_I \rho_I^T + \rho_I u_I^T + u_I \rho_I^T + \sigma^2(n^{-1} Z_I Z_I^T - I),$$

so that

$$\|E_I\|_2 \leq v_s \|\rho_I\|_2^2 + 2\|\rho_I\|_2 \|u_I\|_2 + \sigma^2[\lambda_{max}(n^{-1} Z_I Z_I^T) - 1].$$

Consider the first term on the right side. Since $\|\rho_I - \rho\|_2 \stackrel{a.s.}{\to} 0$ from step (ii), it follows that $\|\rho_I\|_2 \stackrel{a.s.}{\to} \|\rho\|$. As before $v_s \stackrel{a.s.}{\to} 0$, and so the first term is asymptotically negligible.

Let $Z_{I^+} = (z_{\nu i} \ : \ \nu \in I_n^+, i = 1, \ldots, n)$ and $u_{I^+} = (u_\nu \ : \ \nu \in I_n^+)$. On the event $\Omega_n = \{\hat{I} \subset I_n^+\}$, we have

$$\|u_I\| \leq \|u_{I^+}\|$$

and setting $k_+ = |I_n^+|$, by the same arguments as led to (40), we have

$$\|u_{I^+}\|^2 \stackrel{\mathcal{D}}{=} \sigma^2(k_+/n)(\chi^2_{(n)}/n)(\chi^2_{(k_+)}/k_+) \stackrel{a.s.}{\to} 0,$$

since $k_+ = o(n)$ from step (i).

Finally, since on the event Ω_n, the matrix Z_{I^+} contains Z_I, along with some additional rows, it follows that

$$\lambda_{max}(n^{-1}Z_I Z_I^T - I) \leq \lambda_{max}(n^{-1}Z_{I^+} Z_{I^+}^T - I) \stackrel{a.s.}{\to} 0$$

by (38), again since $k_+ = o(n)$. Combining the previous bounds, we conclude that $\|E_I\|_2 \to 0$.

The separation $\delta_n = \|\rho_I\|_2^2 \to \|\rho\|_2^2 > 0$ and so by the perturbation bound

$$\text{dist}(\hat{\rho}_I, \rho_I) \leq (4/\delta_n)\|E_I\|_2 \stackrel{a.s.}{\to} 0.$$

Acknowledgements. The authors are grateful for helpful comments from Debashis Paul and the participants at the Functional Data Analysis meeting at Gainesville, FL. January 9-11, 2003. This work was supported in part by grants NSF DMS 0072661 and NIH EB R01 EB001988.

References

Dembo, A. & Zeitouni, O. (1993), *Large Deviations Techniques and Applications*, Jones and Bartlett, Boston, London.

Donoho, D. (1993), 'Unconditional bases are optimal bases for data compression and statistical estimation', *Applied and Computational Harmonic Analysis* **1**, 100–115.

Geman, S. (1980), 'A limit theorem for the norm of random matrices', *Annals of Probability* **8**, 252–261.

Golub, G. H. & Van Loan, C. F. (1996), *Matrix Computations*, 3rd edn, Johns Hopkins University Press.

Hampton, J. R. (1997), *The ECG made Easy*, Churchill Livingston.

Johnstone, I. M. (2001), Chi square oracle inequalities, *in* M. de Gunst, C. Klaassen & A. van der Waart, eds, 'Festschrift for Willem R. van Zwet', Vol. 36 of *IMS Lecture Notes - Monographs*, Institute of Mathematical Statistics, pp. 399–418.

Johnstone, I. M. (2002), Threshold selection in transform shrinkage, *in* E. D. Feigelson & G. J. Babu, eds, 'Statistical Challenges in Modern Astronomy, III', Springer Verlag, New York. to appear.

Lu, A. Y. (2002), Sparse Principal Component Analysis for Functional Data, PhD thesis, Stanford University, Dept. of Statistics.

Muirhead, R. J. (1982), *Aspects of Multivariate Statistical Theory*, Wiley.

Pratt, J. W. (1960), 'On interchanging limits and integrals', *Annals of Mathematical Statistics* **31**, 74–77.

Ramsay, J. O. & Silverman, B. W. (1997), *Functional Data Analysis*, Springer.

Rice, J. A. & Silverman, B. W. (1991), 'Estimating the mean and covariance structure nonparametrically when the data are curves', *Journal of the Royal Statistical Society, Series B (Methodological)* **53**, 233–243.

Silverman, B. W. (1996), 'Smoothed functional principal components analysis by choice of norm', *Annals of Statistics* **24**(1), 1–24.

Silverstein, J. W. (1985), 'The smallest eigenvalue of a large dimensional wishart matrix', *The Annals of Probability* **13**, 1364–1368.

Stewart, G. W. & Sun, J.-g. (1990), *Matrix Perturbation Theory*, Academic Press.